OXFORD

PhaTYP: predicting the lifestyle for bacteriophages using BERT

Jiayu Shang, Xubo Tang and Yanni Sun

Corresponding author: Yanni Sun, Department of Electrical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China SAR.
Tel: +852-34427612; E-mail: yannisun@cityu.edu.hk

Abstract

Bacteriophages (or phages), which infect bacteria, have two distinct lifestyles: virulent and temperate. Predicting the lifestyle of phages helps decipher their interactions with their bacterial hosts, aiding phages' applications in fields such as phage therapy. Because experimental methods for annotating the lifestyle of phages cannot keep pace with the fast accumulation of sequenced phages, computational method for predicting phages' lifestyles has become an attractive alternative. Despite some promising results, computational lifestyle prediction remains difficult because of the limited known annotations and the sheer amount of sequenced phage contigs assembled from metagenomic data. In particular, most of the existing tools cannot precisely predict phages' lifestyles for short contigs. In this work, we develop PhaTYP (Phage TYPe prediction tool) to improve the accuracy of lifestyle prediction on short contigs. We design two different training tasks, self-supervised and fine-tuning tasks, to overcome lifestyle prediction difficulties. We rigorously tested and compared PhaTYP with four state-of-the-art methods: DeePhage, PHACTS, PhagePred and BACPHLIP. The experimental results show that PhaTYP outperforms all these methods and achieves more stable performance on short contigs. In addition, we demonstrated the utility of PhaTYP for analyzing the phage lifestyle on human neonates' gut data. This application shows that PhaTYP is a useful means for studying phages in metagenomic data and helps extend our understanding of microbial communities.

Keywords: phage lifestyle prediction, virulent and temperate phages, BRET, deep learning

Introduction

Bacteriophages (aka phages) are viruses that infect bacteria. They are widely regarded as the most abundant and diverse entities in the biosphere [1] and play an essential role in various ecosystems [2, 3]. For example, by lysing the bacterial host, phages can regulate both the composition and function of the microbiome. With the in-depth study of phages, there is accumulating evidence revealing phages' significant impacts on various fields, such as dairy production [4, 5], phage therapy [6, 7] and disease diagnostics [8, 9].

Phages' applications depend on annotations of their lifestyles. There are two types of phages based on their lifestyles: virulent phages and temperate phages. Virulent phages infect bacteria and kill their hosts to release their offsprings. But they don't integrate their genomes into the hosts [10]. In contrast, temperate phages integrate their genomes into the host chromosome and copy their genomes together with the host [11]. They will maintain this living state, which is also called prophage, until induced by appropriate conditions and enter the lytic cycle to kill their hosts [12, 13]. The lifestyle of the phages can directly affect their usages. For example, virulent phages are required in phage therapy to kill the antibiotic-resistant bacteria [14, 15], whereas temperate phages

can engineer a host's genome [16] and help regulate gene expression and change cell physiology by introducing novel functions [17, 18]. However, culturing and isolating phages in lab for identifying the lifestyles are usually expensive and time-consuming [19, 20], especially for phages infecting anaerobes, such as *Clostridioides difficile* and *Mycobacterium tuberculosis* [21–23]. Metagenomics allows sequencing of uncultured dark matter of the microbial biosphere, which can contain a large number of phages [24]. Being able to annotate the lifestyles of phages sequenced from host-associated or natural environments is expected to extend our knowledge about phage composition and their interactions with other microbes. Thus, computational prediction of the phage lifestyles has become an attractive alternative to experimental methods.

There are two main challenges for computational prediction of the lifestyle of phages. First, the number of reference phages with known lifestyle annotations is very limited. According to the latest lifestyle annotation dataset provided by [25], there are 1290 virulent phages and 577 temperate phages. However, the number of released phages in the RefSeq database is 4517 in 2021, indicating that over a half of phages have no annotations. An even larger data source for phage is IMG/VR v3 database [26], which contains nearly 2 million uncultivated phage-like genomes

---

in 2021. Because of the limited number of annotated genomes, most phages cannot be classified by sequence match-based methods. Second, as mobile genetic elements, phages usually mobilize host genetic material and incorporate it into their own genomes [27], leading to poor sequence assembly results and incomplete fragments for phages [25]. Both the short length of the fragments and the ambiguous regions increase the difficulty of the lifestyle prediction task.

Related work

Given the importance of phages, numerous efforts have been made to computationally predict phages' taxonomic labels [28–30], to identify their hosts [31–34] and to detect phages from metagenomic data [35, 36]. Nami *et al.* reviewed some of these tools recently [37]. A handful of these studies focus on phage lifestyle prediction [19, 25, 38]. One type of method uses maker genes to distinguish virulent and temperate phages. For example, integrase and excisionase are two widely accepted marker genes for identifying temperate phages [39]. However, only a few genes can be used as marker genes, especially for virulent phages [19]. In addition, the fragmented contigs from the metagenomic assembly may not cover such genes, and thus, using a small set of marker genes can lead to a low recall for lifestyle classification.

Instead of relying on handcrafted marker genes, learning-based methods are proposed aiming to automatically learn features from two types of phages' DNA and protein sequences. For example, PHACTS [19] trained a random forest model on protein similarities for virulent and temperate phage classification. BACPHLIP [38] trained another random forest classifier using a set of lysogeny-associated protein domains identified by HMMER3 [40]. However, such strategies may not apply to metagenomic data. According to the benchmark results shown in [25], the accuracy of PHACTS decreases on short contigs. For example, PHACTS only achieves an accuracy of ˜60% when the phage contigs are below 2 kbp. Also, BACPHLIP is only designed for complete phage genomes according to the provided guidelines; it returns errors when short contigs are provided as inputs. Unlike these methods, PhagePred [41] and DeePhage [25] can identify the lifestyle for contigs assembled from metagenomic data. PhagePred calculated the distance between a query and a reference using a learned *k*-mer frequency-based Markov model. DeePhage [25], which has the best reported performance on lifestyle prediction, can predict contigs as short as 100 bp. DeePhage distinguishes the lifestyle of phages by applying a convolutional neural network to learn the motif-related information from DNA sequences. Nevertheless, its best performance on the short contigs is only ˜80%.

Overview

In this work, we present a method named PhaTYP to classify the lifestyles of phages. Previous works have shown that the marker genes and protein-protein associations are key features in phage classification and identification [28–30, 34, 35, 42]. The previous studies also showed that the protein composition and their associations play important roles on phages' lifestyle [39, 43]. Inspired by these studies, we represent phage contigs using a contextualized embedding model from natural language processing (NLP) for lifestyle classification. Specifically, we adopt Bidirectional Encoder Representations from Transformer (BERT) to learn the protein composition and associations from phage genomes. We evaluated our final model PhaTYP on contigs of different lengths and contigs assembled from real metagenomic data. The benchmark results against the state-of-the-art methods

show that PhaTYP not only achieves the highest performance on complete genomes but also improves the accuracy on short contigs by over 10%.

Methods

Transformer has emerged as a powerful general-purpose model architecture for representation learning. Because the multi-head mechanism implemented in transformer can learn the association between tokens, transformer can be used to generate the semantic representation of the sentences. It outperforms recurrent and convolutional neural networks in several NLP tasks. Inspired by the usage of transformer in NLP, we propose PhaTYP, which is an eight-layer bidirectional transformer block based on the original implementation described in [44]. In NLP problems, words are the tokens in sentences. We make an analogy between words in NLP and proteins in phage genomes so that we can utilize transformer to learn protein composition and associations from phage genomes. Although k-mers and motifs have been used as tokens for protein prediction tasks [45, 46], using proteins as tokens can integrate their biological functions into the sentence representation. In addition, the converted sentence is much shorter by using protein-based tokens, making the training much faster with less parameters. Thus, we train PhaTYP on protein-based tokens to separate virulent and temperate phages.

To address the difficulties of classifying incomplete genomes with limited training data, we divide the lifestyle classification into two tasks: a self-supervised learning task (Figure 1 A) and a fine-tuning task (Figure 1 B). In the first task, to circumvent the problem that only a limited number of phages have lifestyle annotations, we applied self-supervised learning to learn protein association features from all the phage genomes using Masked Language Model (Masked LM), aiming to recover the original protein from the masked protein sentences. This task allows us to utilize all the phage genomes for training regardless of available lifestyle annotations. In the second task, we will fine-tune the Masked LM on phages with known lifestyle annotations for classification. To ensure that the model can handle short contigs, we apply data augmentation by generating fragments ranging from 100 to 10 000bp for training.

In the following section, we will first describe how to convert DNA sequences into protein-based sentences. Then, we will briefly introduce the background of eight-layer bidirectional transformer blocks. Because we adopt a very standard implementation of transformer, we refer readers to [44] for detailed explanations of the model. Finally, we will show how we train PhaTYP on two different tasks: self-supervised training for Masked LM and the fine-tuning model for lifestyle classification.

Sequence embedding

In order to convert sequences into protein-based sentences, we will first introduce how we construct the token vocabulary. Each token in our model represents a protein cluster containing homologous protein sequences from phages. First, we downloaded all the phages proteins from the RefSeq database. We run an all-against-all similarity search using DIAMOND BLASTP [47] and generate a protein-similarity graph, where the nodes in the graph represent the protein, and the edges connect proteins with significant sequence similarities. The edge weights are the e-values returned from DIAMOND BLASTP. Then, we employ Markov clustering algorithm [48] to group proteins into clusters based on their similarity (e-value). This process resulted in 63 855 protein clusters. Finally, we remove all clusters that contain only one protein
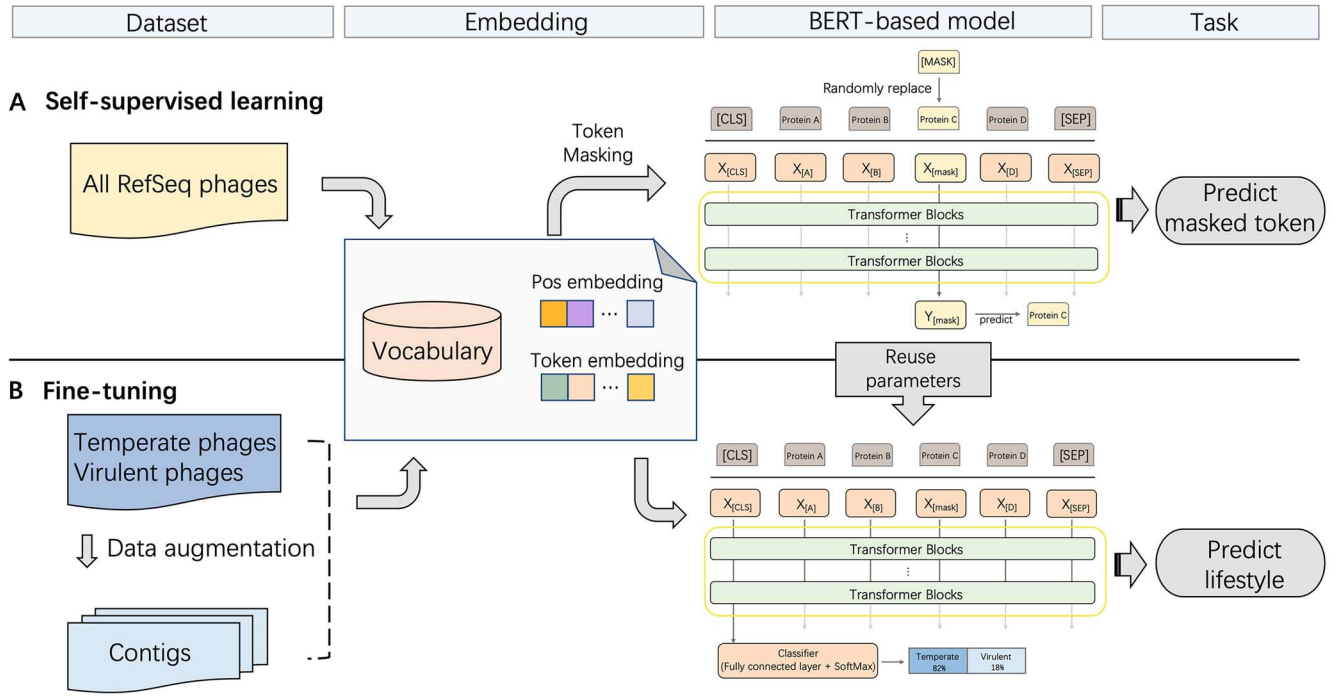
Figure 1. Two training tasks for PhaTYP using BERT. (**A**) The self-supervised learning task. The input is the masked sentence, and the output is the predicted token at the masked position.'1 All phage genomes in RefSeq database to train a Mask LM model. (**B**) The fine-tuning task for lifestyle prediction. The pretrained model is fine-tuned using phages with known lifestyle annotations. The inputs of the model are protein-based sentences, and the outputs are the probabilities of two lifestyle classes: virulent and temperate.

and use the remaining 45 578 protein clusters for constructing the token vocabulary.

We then use the generated vocabulary to convert contigs into protein-based sentences. As shown in Figure 2, first, Prodigal [49] is adopted for gene finding and translation. Then, DIAMOND BLASTP is applied to measure the similarity between these query proteins and the protein clusters. Each query protein is assigned with the token with the best alignment. Finally, the contigs can be converted into protein-based sentences as shown in Figure 2. Because the length of the sentences can vary a lot, we follow the design of BERT [50] and choose 300 as the maximum length of the sentences. Thus, the sentence is a 300-dimensional vector, with each dimension encoding a token ID. Two specialized tokens, [*CLS*] and [*SEP*], representing the start of the sentences and separation of the sentences, are the first and last tokens in each sentence. If the contigs have more than 298 tokens, we only keep the first 298 tokens. On the contrary, if the contigs contain less than 298 tokens, we pad token [*PAD*] at the end of the sentences. In the RefSeq database, most of the phages (95.9%) contain less than 298 tokens. For a small percentage of sequences containing more than 298 tokens, using the first 298 or the last 298 tokens leads to almost identical performance based on our empirical studies.

After converting the DNA sequences into protein-based sentences, we will project the 300-dimensional vector into a dense embedding matrix. We employ a learnable embedding layer, which is a neural network, to embed the sentence. There are two main purposes of using the learnable embedding layer. First, the learnable embedding layer will generate a low-dimensional embedding vector than using one-hot encoding as input, which can result in resulting in a $\mathbb{R}^{300 \times 45,578}$ matrix for each sentence. The deep learning model will suffer from the curse of dimensionality using such a sparse matrix as input [51]. Second, as proven in [51], the embedding layer can learn to map associated tokens
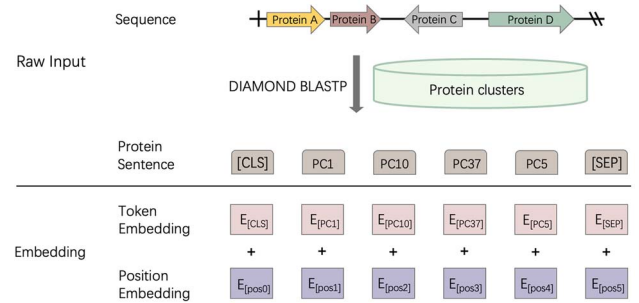

Figure 2. Sequence embedding method in PhaTYP. The block in 'Protein Sentence' represents the ID of the protein-based token. PCx: a protein cluster x. [CLS]: start token. [SEP]: separation token. $E_{[PCx]}$: the embedded vector for protein cluster PCx. '+': vector addition.

into similar embedding vectors, and thus assisting the learning process. In addition, as shown in Figure 2, we embed the position information to represent the position of the token in the sentence, helping the model utilize the sequential information of the sentence. The equations of the embedding layers are listed in Equation 1.

$$\begin{cases} \widetilde{E}_t = Embed(E_t, W_{E_t}) \\ \widetilde{E}_p = Embed(E_p, W_{E_p}) \\ \quad X = \widetilde{E}_t + \widetilde{E}_p \end{cases} \quad (1)$$

$E_t \in \mathbb{R}^{300 \times 1}$ represents the token ID sentence and $E_p \in \mathbb{R}^{300 \times 1}$ represents the position index vector. $W_{E_t} \in \mathbb{R}^{45,578 \times 512}$ and $W_{E_p} \in \mathbb{R}^{300 \times 512}$ are the learnable projection matrices. 512 is the default embedding dimension suggested by [44]. The function of *Embed* is a look-up table. Given an ID, it returns the corresponding
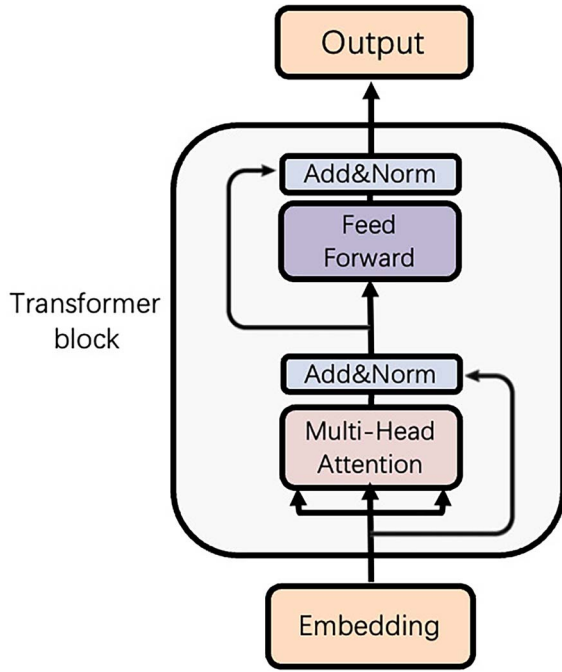
Figure 3. The transformer block in PhaTYP. There are three main units in the transformer block: feed-forward network, residual connections and multi-head attention mechanism.

embedding vector. Thus, the dimension of $\widetilde{E}_t$ and $\widetilde{E}_p$ are $\mathbb{R}^{300\times512}$, which is much smaller than using the one-hot encoding matrix ($\mathbb{R}^{300\times45,578}$). We follow the design of BERT [50] and train the embedding layer with the whole model via the end-to-end mode to enlarge the receptive field. Finally, we will apply matrix addition on the two embedded matrices (defined as X in Equation 1) and feed them into the eight-layer transformer structure.

Model structure

The detailed architecture of the transformer model is shown in Figure 3. The main function is the multi-head attention mechanism, which can extract association features between tokens [44]. The feed-forward network consists of two fully connected layers with a ReLU activation, which is applied to each position separately and identically. To prevent overfitting, residual connections [52], and layer normalization [53] are employed before and after the feed-forward network. The input of the transformer block is the embedding X in Equation 1. The self-attention mechanism and feed-forward network of the transformer grant access to all protein tokens in the embedded sentence and leverage the association information between tokens to generate latent feature output Y. Because of the hyper-parameter setting in the transformer model, Y has the same dimension as the embedding X. In PhaTYP, we stack eight transformers in serial. In each transformer, we employ eight-head attention to focus on different representation sub-spaces.

After introducing the model structure, we will show how to train PhaTYP on Masked LM and fine-tuning tasks. These two tasks have different but related objective functions. In the following sections, we will first describe the datasets used in each task and then present a detailed explanation of the loss function.

Self-supervised training for Masked LM

Self-supervised learning is a kind of unsupervised learning aiming to learn a general feature expression for downstream tasks. In our design, the downstream task is the lifestyle classification, and the general feature is the protein organizations in phage genomes. Because the number of known lifestyle annotation of phages is far less than the number of known phages in the database, we employ a self-supervised learning method to learn as much prior knowledge as possible. Specifically, we use all the available phage genomes in RefSeq to train the Masked LM task.

Dataset. We download all the phage genomes released before 2022 from the NCBI RefSeq database. In total, we have 3474 phage genomes. To generate more data for training, we cut the complete genomes into fragments with different lengths, including 5, 10, 15 and 20kbp. We randomly sample 10 sub-strings from the genomes for each length. Thus, we have 142 434 phage contigs in our dataset. Finally, we will use the method introduced in section *sequence embedding* to generate protein sentences and embedding as inputs to transformer.

Masked LM loss. The inputs to our self-supervised model are masked sentences, and the aim is to predict the original sentences. For example, as shown in Figure 1A, we replace the embedding at the position of *protein C* with the [*MASK*] token. Then, we will feed the masked sentence into PhaTYP to predict the marked token.

As described in Section Model structure, the shape of output Y will be the same as the input X, which is $\mathbb{R}^{300\times512}$. Thus, each row i in Y can be interpreted as the latent features of the input token at position i in the sentence, which is a $\mathbb{R}^{1\times512}$ vector. Then, in order to predict the ID of the masked token, Y_i is fed into an output layer with SoftMax function over the vocabulary. The loss function is presented in Equation 2.

$$loss(SoftMax(Y_iW_d), Token_i) \qquad (2)$$

$W_d \in \mathbb{R}^{512\times45,578}$ is the learnable projections. The value in $SoftMax(Y_iW_d) \in \mathbb{R}^{1\times45,578}$ represents the prediction probability of each token in the vocabulary. In the training process, 5% of words of each sentence will be masked randomly, and we will calculate the cross-entropy loss to update the parameters through back-propagation.

Fine-tuning the model for lifestyle classification

After pretraining the model via the self-supervised learning task, we will then fine-tune PhaTYP to the downstream task: lifestyle classification.

Dataset. There are two widely used datasets supplied by the previous studies [19, 25]. In total, we have 1290 virulent phages and 577 temperate phages. In order to balance the dataset and improve the robustness on short contigs, data augmentation is applied by randomly generating short DNA fragments, ranging from 100bp to 20kbp, from the complete genomes. For each length, we generate 10 000 contigs for temperate and virulent phages, resulting in 160 000 phage contigs in the dataset. Then, all the fragments will be converted into protein sentences and fed into PhaTYP.

Fine-tuning loss. Unlike the self-supervised learning task, we use the first row $Y_0 \in \mathbb{R}^{1\times512}$ from output Y as latent feature for the lifestyle classification. As shown in Figure 1B, we concatenate a fully connected layer to calculate the probability of the input being virulent or temperate. The loss function of the fine-tuning can be found in Equation 3

$$loss(SoftMax(Y_0W_p + b_p), label) \qquad (3)$$

Table 1. Detailed information of the dataset used in the experiments

Name	Description
RefSeq	All the phage genomes released before 2022 from the NCBI RefSeq database. Totally, we have 3474 phage genomes. This dataset is only used to train the Masked LM task.
Lifestyle dataset 1	Lifestyle annotations from the dataset of [19], including 77 virulent phages and 148 temperate phages.
Lifestyle dataset 2	Lifestyle annotations from the dataset of [25], including 1211 virulent phages and 429 temperate phages.
Phage contigs from human neonate's guts	2291 phage contigs assembled from metagenomic data. The metagenomic data is sampled from 20 human neonate's guts [54].

$W_d \in \mathbb{R}^{512 \times 2}$ and $b_p \in \mathbb{R}^{1 \times 2}$ are the learnable parameters. Then, we will calculate the loss between the real labels and the predictions and update the parameters in the model accordingly.

Model training

In the training process, we employ 10-fold cross-validation for both tasks. The model is trained with 4 GTX 2080Ti GPU units using the Adam optimizer, and we apply a learning rate of 0.001 to update the parameters. Because PhaTYP needs to be trained step-by-step on the two tasks, we first choose the model with the lowest cross-validation loss on self-supervised learning tasks. Then, we fine-tune this pretrained model on the lifestyle classification task. Finally, we store the parameters of the model that achieve the best performance (AUCROC) in the cross-validation. Both the parameters and the model are available via http://github.com/KennthShang/PhaTYP/.

Results

In this section, we validate PhaTYPE on both simulated and real sequencing data. We compared PhaTYP against the state-of-the-art methods, including PHACTS [19], DeePhage [25], BACPHLIP [38] and PhagePred [41]. Because the authors of BACPHLIP stated that incomplete/partially assembled genomes should not be used as input, we will only evaluate BACPHLIP on complete genomes. It is worth noting that PhagePred and DeePhage allow re-training using customized training data. Thus, we re-trained these models with the same training set as PhaTYP for all experiments. PHACTS and BACPHLIP did not provide this function, and thus, we used the provided model in all experiments.

Metrics

To ensure consistency and a fair comparison, we use the same metrics as the previous works: sensitivity [$TP/(TP+FN)$], specificity [$TN/(TN + FP)$] and accuracy [$(TP + TN)/(TP + FN + TN + FP)$]. TP represents the number of correctly classified temperate phages, whereas TN represents the number of correctly classified virulent phages. Thus, sensitivity and specificity can be interpreted as the recall of classifying temperate and virulent viruses, respectively. In addition, we will present the ROC curve to evaluate the trade-offs between sensitivity and specificity.

Datasets

As described in Section *Self-supervised training for Masked LM* and *Fine-tuning the model for lifestyle classification*, totally we have three datasets for training and validation. In addition, we applied PhaTYP to predict the lifestyle for phage contigs assembled from human neonates' guts. The detailed information of the datasets is listed in Table 1.
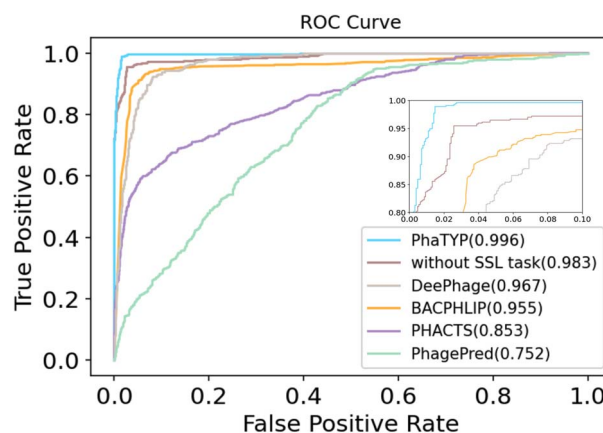
Figure 4. The ROC curve comparison on the complete phage genomes. The value shown in the legend is the AUCROC score. 'Without SSL task': training PhaTYP without the self-supervised learning task.

Classification performance comparison using 10-fold cross-validation

Performance on the complete genomes. We applied ten-fold cross-validation on the combined lifestyle datasets listed in Table 1. When training, we apply the data augmentation method mentioned in section *Fine-tuning the model for lifestyle classification* on the training set. The phage sequences in the validation set remain complete.

The ROC curves derived from the averaged ten folds evaluation from all the methods are shown in Figure 4. We also show how the self-supervised learning task affects the learning performance by recording the results of training PhaTYP without the self-supervised learning task (*without SSL task* in Figure 4). Predicting the lifestyle for complete phages genomes is a relatively easy task. PhaTYP, DeePhage and BACPHLIP all achieved high accuracy. Nevertheless, the ROC curves in Figure 4 show that PhaTYP has the best performance.

In addition, we listed more detailed information about the performance of predicting virulent and temperate phages in Table 2. The results reveal that PhaTYP can identify temperate phages with higher specificity than other tools. The ablation study also shows that training with the self-supervised learning task can improve classification accuracy.

We also recorded the running time of the five methods on the 10-fold cross-validation in Figure 5. PhaTYP is not the fastest tool because ~85% of running time is used to run DIAMOND BLASTP as described in Section *Sequence embedding*. DeePhage requires less running time because it only uses k-mer features, whereas other methods also incorporate alignment features for lifestyle prediction.

Performance on the test set with low similairty. Cross-validation cannot control the similarity between the training and

Table 2. Detailed results on the complete genomes under each tools' default score cutoffs (0.5)

| Metrics	Sensitivity	Specificity	Acc
PhaTYP	**0.99**	**0.97**	**0.98**
Without SSL task	0.97	0.91	0.95
DeePhage	0.94	0.93	0.94
BACPHLIP	0.97	0.87	0.93
PHACTS	0.91	0.75	0.85
PhagePred	0.69	0.88	0.81
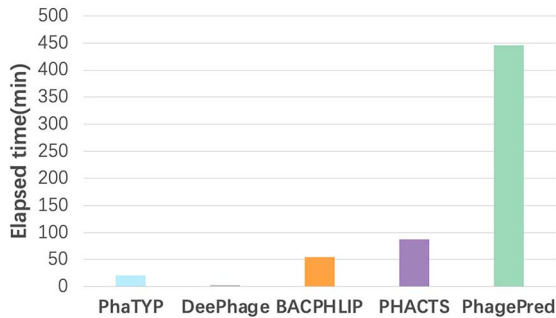
Figure 5. The running time comparison of five tools in the experiment of ten-fold cross validation. All the methods are run on Intel® Xeon® Gold 6258R CPU and 2080Ti GPU.

Table 3. The performance comparison on the low similarity test set under the tools' default score cutoff (0.5)

Metrics	Sensitivity	Specificity	Acc
PhaTYP	**0.99**	**0.89**	**0.94**
Without SSL task	0.98	0.86	0.92
DeePhage	0.96	0.86	0.91
BACPHLIP	0.98	0.84	0.90
PHACTS	0.90	0.69	0.74
PhagePred	0.57	0.83	0.67

test set. To evaluate whether the learning models can perform well on a harder test set, we constructed a test set containing phage genomes with low similarity with the training set. First, we applied Dashing [55] to estimate the similarity between phages in the lifestyle dataset. Then, we generated a test set by selecting 50 phages with the lowest similarities with their peers from virulent and temperate phages, respectively. In total, there are 100 phages in the test set. The maximum similarity of the virulent phages and temperate phages in the test set with all other phages are 0.049 and 0.052, respectively.

The classification results are shown in Table 3 and Figure 6. Compared with the cross-validation results in Table 2, the similarity of the test genomes affects specificity more than sensitivity, indicating that some virulent phages are misclassified as temperate. This could be caused by lower average similarity for the test virulent phages. In Figure 6, the value of AUCROC decreases for all methods except BACPHLIP. This is likely caused by the overlap between the test phages and the data used for training BACPHLIP in its provided version. Nevertheless, PhaTYP still has the best results.

Performance on the short contigs. Considering that metagenomic assembly can produce incomplete phage sequences, it is important to evaluate PhaTYP on phage contigs. Toward this goal,
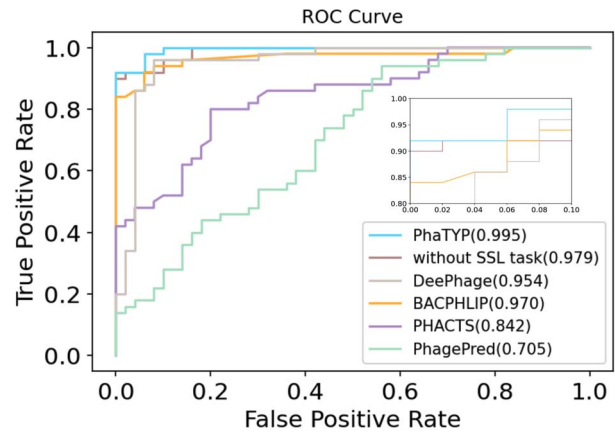
Figure 6. The ROC curve comparison on low similarity test set. The value shown in the legend is AUCROC score. 'without SSL task': training PhaTYP without self-supervised learning task.
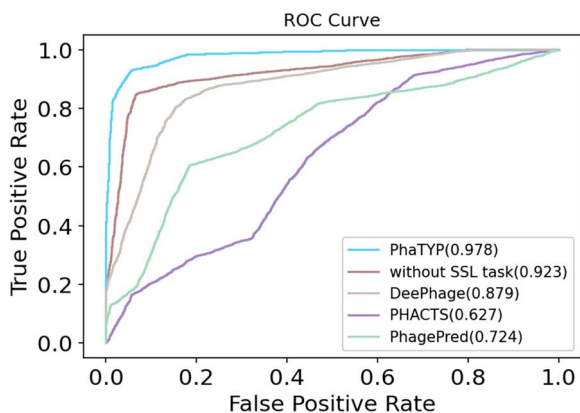
we apply the same method used in [25] to construct the short fragment dataset. Specifically, we generate four groups of fragments with different length ranges, including 100–400, 400–800, 800–1200 and 1200–1800 bp. These four sets of fragments can cover the length of raw reads and the short contigs from the metagenomic data. We generated 80 000 fragments for each group by random sampling a sub-string from the complete genomes, with 40 000 for temperate and virulent phages, respectively. To remove the potential redundant fragments, we used Dashing [55] to estimate the similarity between fragments. Then, we removed fragments with similarities above 0.8 from the dataset. We applied ten-fold cross-validation to train and validate the performance. Because DeePhage trained four separate models on different length ranges, we followed the same design and trained four PhaTYP models separately.

We generate the ROC curve on short contigs to show the tradeoff between FP rate and sensitivity. In each fold, we combine the validation results from the 4-length ranges. The results are shown in Figure 7, which reveals that our model outperforms the other tools. More detailed results are shown in Table 4 and Figure S1 in the Supplementary File. In general, with the increase of the length, the performance of all methods increases. This is expected because longer fragments generally provide more information for prediction. In addition, compared with the results on complete genomes, the self-supervised learning task can help the model achieve higher accuracy on short contigs. A plausible explanation is that pre-training on the self-supervised learning task can help the model learn more generalized embeddings for phage proteins and prevent overfitting. Then, these embeddings with prior knowledge can be leveraged when information for classification is lacking. We also calculated the P-value of the ROC curve between PhaTYP and the state-of-the-art methods. Table S1 in the Supplementary File shows that all the differences are statistically significant, indicating that PhaTYP can achieve better performance in classifying phages' lifestyles.

In addition, we tested whether training contigs with different lengths at the same time influences the performance of PhaTYP. Specifically, we employed the data augmentation methods, which combines the training set of the contigs and their complete genomes to train one PhaTYP model. The comparison between training PhaTYP separately and training with all data at once (PhaTYP augmentation) in Table 4 reveals that training with data augmentation will not affect the classification performance.

Table 4. Detailed results on the short contigs under the tools' default score cutoff (0.5)

Tool	Criterion	100–400	400–800	800–1200	1200–1800
PhaTYP	Sensitivity	**0.86**	0.91	**0.93**	0.93
	Specificity	**0.92**	0.92	**0.93**	**0.94**
	Accuracy	**0.89**	0.91	**0.93**	**0.94**
without	Sensitivity	0.83	0.85	0.87	0.92
SLL task	Specificity	0.92	**0.93**	0.92	0.93
	Accuracy	0.87	0.89	0.90	0.92
DeePhage	Sensitivity	0.74	0.84	0.86	0.89
	Specificity	0.77	0.82	0.86	0.87
	Accuracy	0.76	0.83	0.86	0.88
PhagePred	Sensitivity	0.56	0.60	0.62	0.64
	Specificity	0.75	0.80	0.84	0.87
	Accuracy	0.65	0.70	0.72	0.75
PHACTS	Sensitivity	0.26	0.36	0.39	0.42
	Specificity	0.73	0.64	0.65	0.69
	Accuracy	0.48	0.49	0.51	0.54
PhaTYP	Sensitivity	**0.86**	**0.92**	0.92	**0.94**
(augmentation)	Specificity	0.90	0.92	**0.93**	**0.94**
	Accuracy	0.88	**0.92**	**0.93**	**0.94**

Figure 8. The classification results on the 33 virulent crAssphages. PhaTYP can correctly predict all the crAssphages as being virulent.
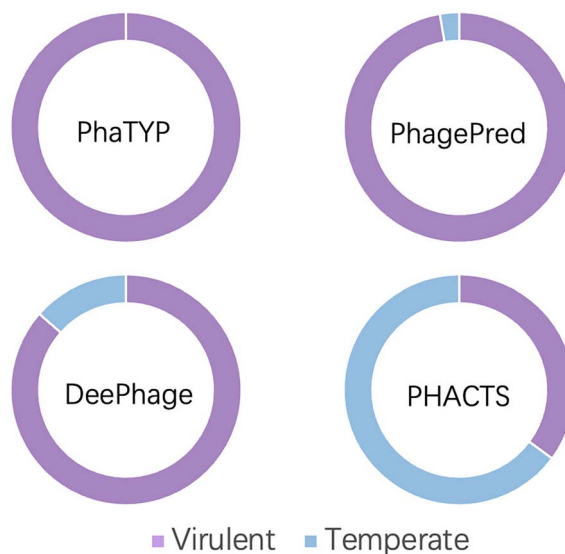

Figure 7. The ROC curve comparison on the short contigs. The value shown in the legend is AUCROC score. 'without SSL task': training PhaTYP without self-supervised learning task. PhaTYP has the best performance.

Thus, PhaTYP can predict lifestyles for contigs across different length ranges with one model, which helps save computational resources for users.

Predicting the lifestyle of crAssphages

CrAssphages are an extensive family of tailed bacteriophages discovered through the cross-assembly of human fecal metagenomes [56]. Most of them infect bacteroides and do not integrate into their hosts during replication. Thus, these crAssphages are widely regarded as virulent phages. Despite its ubiquity in human gut, over 80% of the predicted proteins in crAssphage genomes showed no significant similarity to annotated crAssphage protein sequences, hampering their identifications in newly sequenced metagenomic data [57]. The low similarity also poses a hard case for current lifestyle classification tools.

In this experiment, we downloaded 33 crAssphages from [54], which are all annotated as virulent phages, and evaluated whether PhaTYP can correctly classify them. First, we remove all the training sequences with high sequence similarity (alignment identity > 50%) from our dataset using BLASTN [58]. Then, we

retrained PhaTYP, DeePhage and PhagePred on the 'cleaned' dataset.

The results of this experiment are presented in Figure 7. All the 33 crAssphages can be classified correctly by PhaTYP, further demonstrating its utility. PhagePred ranked second, showing much improved performance than in the previous experiments. One possible reason is that the k-mer frequency feature adopted by PhagePred is an important feature for crAssphages.

Case study: phage lifestyle analysis for infants

In this section, we apply PhaTYP to investigate early-life viral colonization using the metagenomic data sampled from infant meconium or early stools [54]. In the original study, the authors used reference genomes and PHACTS to analyze the phages' lifestyles. Although the similarity search against the reference genomes can be used to annotate the lifestyles, the number of aligned phages is only a small subset of all assembled phage contigs. Then, the authors used PHACTS to predict lifestyle of all the phage contigs. However, according to the previous experiments, the performance of PHACTS is not ideal on short contigs. Thus, we employed PhaTYP to re-investigate this dataset.

The data are sequenced from 20 infants. The stool samples were collected longitudinally at 0–4 days after birth (meconium samples, *month 0*), *month 1* and *month 4*. Therefore, we have a total of 60 metagenomic samples. In addition, the authors recorded detailed information, including the feeding and delivery type of these infants. Thus, we are able to investigate whether the ages, feeding and delivery type can affect the composition of phages with different lifestyles. The 60 samples containing the raw reads are public and available via https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA524703; the assembled contigs are available via https://github.com/guanxiangliang/liang2019.

Following the guidelines of [54], we first applied Bowtie2 [59] for reads mapping and then ran BBmap [60] to calculate the reads per million total reads (RPM) for each contig. According to the description in [54], the species were called present if at least 10 RPM from one sample aligned to that contig. Therefore, we have 19 943 contigs remained. Second, to ensure consistency, we used the same rules to define phage contigs as in [60]: (1) at least one
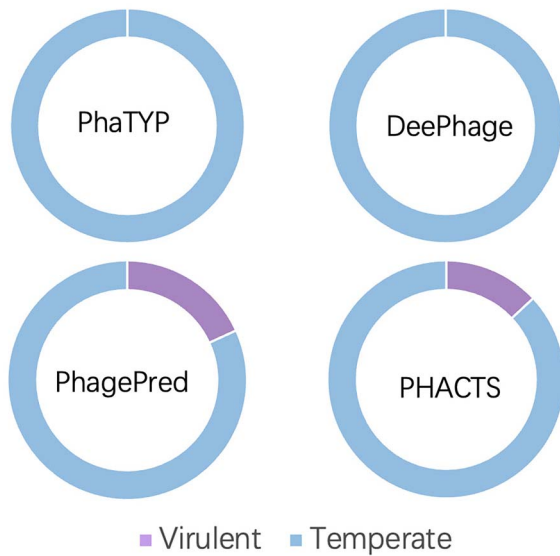
Figure 9. The classification results on the contigs that have homologous regions with integrase proteins. Both PhaTYP and DeePhage can correctly predict all the contigs as being temperate.

Figure 10. Violin plot at different months. (A) Predictions on all 2291 phages. (B) Predictions on newly colonized phages. Y-axis: the percentage of temperate phages in each sample.

phage protein per 10 kb of the contig and (2) 50% of the predicted open-reading frames (ORFs) are phage ORFs. We used the Prodigal [49] in meta mode to predict the ORFs and aligned them to the phage protein database downloaded from RefSeq. Finally, 2291 high-confidence phage-like contigs were identified and we ran PhaTYP for lifestyle prediction.

Identifying temperate phages containing integrase. Because there are no labels for the contigs from the metagenomic data, we rely on the marker genes for labeling the contigs. As we discussed in section *Introduction*, using marker genes usually has high specificity but low recall. Thus, the contigs labeled using this method have high confidence. One of the most widely accepted markers, named integrase, is commonly used to identify temperate phages [61, 62]. Thus, we built an integrase protein database by searching the proteins with the keyword 'integrase' from the phage protein database. Then, we search all 2291 phage contigs against the integrase database using BLASTX [58]. 23 contigs having homologous regions (identity > 90%, coverage > 90% and e-value < 1e-10) were identified out of 2291 contigs. All these 23 contigs were labeled as temperate phages, and we used them to evaluate PhaTYP and the state-of-the-art tools.

As shown in Figure 9, both PhaTYP and DeePhage can predict correctly on all 23 contigs. On the other hand, the results show that using the integrase as a marker can only classify ˜1% of the metagenomic phages, which remains a low recall for lifestyle prediction. Thus, methods that do not just rely on integrase proteins are in great need of comprehensive prediction.

Comprehensive lifestyle analysis. Then, we analyzed all 2291 phages' lifestyles using PhaTYP. There are three main variables that might affect the dominant lifestyle: the *age* of the infant, the *delivery* type and the *feeding* type.

Ages. First, we group the samples by age and draw violin plots to show the lifestyle distribution in Figure 10A. Y-axis represents the percentage of temperate phages in each sample: *temperate*/(*temperate*+*virulent*). We also record the P-value to show whether the two groups differ significantly. Figure 10A shows that as the baby grows, the percentage of temperate phages decreases, suggesting that more virulent phages colonize in the infants' gut
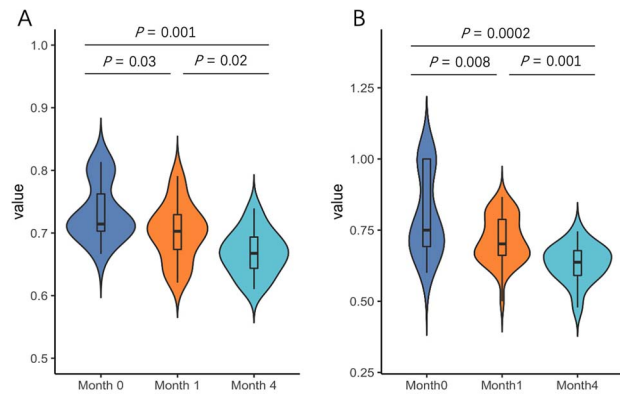
with the infant's growth. Thus, we further investigate the lifestyle of newly colonized phages for each infant. Specifically, because we have three data samples representing *month 0*, *month 1* and *month 4*, we can first identify new phage contigs in month 1 by aligning reads of month 0 to the contigs in month 1 via Bowtie2 [59]. If a contig in month 1 has no read mapping outputs, this contig is regarded as a newly colonized phage after month 0. Finally, the lifestyle prediction results on these newly colonized phages are shown in Figure 10B. The trend and conclusion are the same as Figure 10A and the P-value became smaller, indicating that the newly colonized viruses are more likely to be virulent phages.

Our results are consistent with the main conclusions of the original study [54]. During the early stage after birth, pioneer bacteria colonize the infant's gut. The prophages induced from these bacteria provide the predominant population. Then, as the infant grows, more phages and bacteria from the environment might reside within human guts too, leading to the change of the lifestyle composition.

Delivery type. Because the phage lifestyle on different age groups shows significant differences, we further investigate whether the phage lifestyle composition is influenced by the *delivery* type. We first group the data by age. Then, we draw the violin plots for three delivery types in each group. The results are shown in Figure 11.

There are three types of delivery: *C-Section with labor (CS(w/)L)*, *C-Section without labor (CS(w/o)L)* and *spontaneous vaginal delivery (SVD)*. In group *month 0*, there is a significant difference between the delivery type and phage lifestyle composition. With the increase of the age, the difference gradually becomes smaller, suggesting that delivery type can influence the initial phage colonization of the infants. According to the explanation in [54], the environmental contamination of different delivery types can affect the gut microbiome at *month 0*, which might also lead to the difference.

Feeding type. Because we only have feeding types for infants at *month 1* and *month 4*, we show the two groups' violin plots in Figure 12. An interesting finding is that the feeding type does not show much difference at *month 1*. But with the growth of infants, the proportion of temperate phages in the infants with formula milk decreases more rapidly than those with the mixed feeding type. One possible reason is that breastfeeding can reduce the chance of infection by microbes [63–65] and thus, leading to relatively stable environments for the infants with access to breast milk.
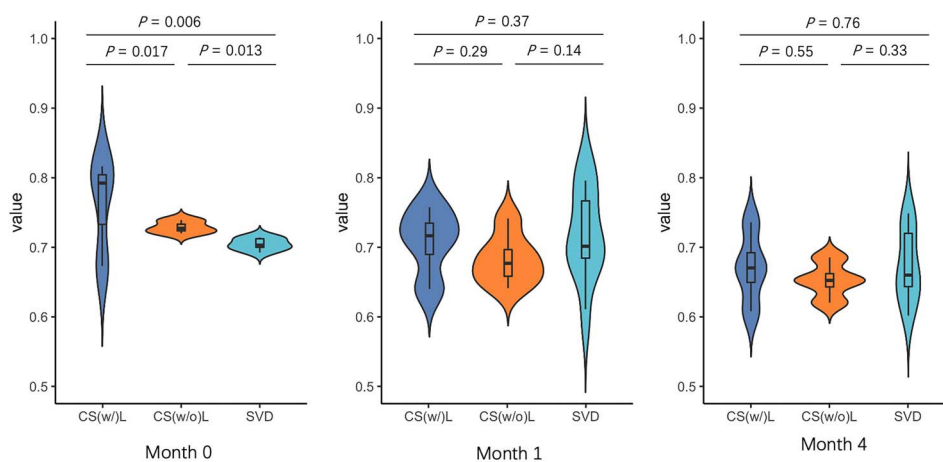
Figure 11. Violin plot in different delivery type: *C-Section with labor (CS(w/)L)*, *C-Section without labor (CS(w/o)L)*, and *SVD*. To control variables, we group the samples according to the ages. The Y-axis represents the percentage of temperate phages in each sample.
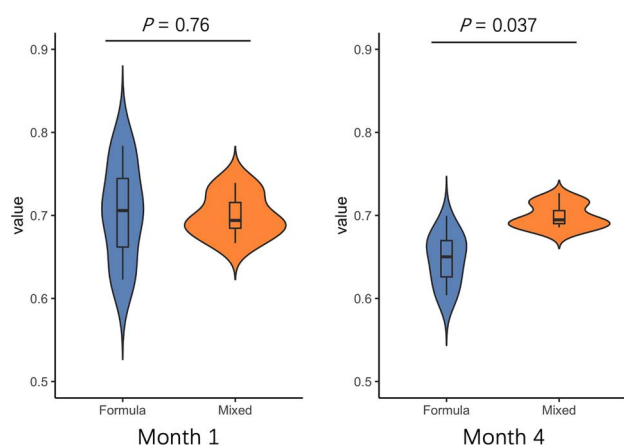
Figure 12. Violin plot in different feeding types. *Formula*: the infants were fed with formula milk. *Mixed*: infants were fed with both formula and breast milk. To control variables, we group the samples according to the ages. The Y-axis represents the percentage of temperate phages in each sample.

Discussion

In this work, we propose a method named PhaTYP for phages' lifestyle prediction. As shown in the experiments, PhaTYP outperforms the available lifestyle prediction methods. We design two different tasks, a self-supervised learning task for learning protein association and a fine-tuning task for lifestyle prediction, to train the model so that PhaTYP achieves more stable performance on short contigs than other methods. In addition, we show an application of using PhaTYP to investigate the early viral colonization in human neonates. The prediction results of PhaTYP yielded several insights, most of which are consistent with recently published studies, suggesting that PhaTYP is practically useful for studying and analyzing phage composition in metagenomic data.

Machine learning has become a powerful means for studying phages. A recent review [37] summarized a number of tools that exploit machine learning models to study various problems about phages, including phage host prediction, phage discovery, taxonomic classification, virion protein identification and life cycle prediction. To our best knowledge, this work is the first to use BERT for phage life style prediction. Because BERT is a powerful LM, it may be extended and customized for other phage-related research problems. For example, the self-supervised learning task can be applied before phage classification and host prediction.

Although PhaTYP has greatly improved phages' lifestyle prediction, we have several aims to optimize PhaTYP in our future work. One possible extension is to employ the knowledge distillation to reduce the parameter size of PhaTYP. Because of the stacking of transformer blocks in PhaTYP, it requires large computational resources (four 24Gb GPU units) to accelerate the training process. Distilling the knowledge in a neural network can provide a light-version model with the same performance for users who wish to re-train the model. Another extension is to integrate other phage analysis programs with PhaTYP. Despite multiple programs for analyzing phages, there lacks a comprehensive pipeline for conducting end-to-end phage characterization for metagenomic data. In our future work, we will incorporate PhaTYP with our previously published tools on phage classification and host prediction [30, 34] and establish a web server for analyzing phage contigs in metagenomic data.

In conclusion, based on our tests on both simulated and real sequencing data, PhaTYP achieves the most accurate lifestyle prediction among available tools. PhaTYP can be applied to various metagenomic data for analyzing virus compositions.

Key Points

- Predicting phages' lifestyles is a key step for understanding the phage population and phage–host interactions in metagenomic data. We developed a lifestyle prediction tool named PhaTYP that employs the BERT model and protein-based tokens to learn the protein composition and their associations in phage sequences.
- To overcome the challenges of classifying short contigs with limited training data, we designed self-supervised and fine-tuning tasks to help the model achieve stable performance on fragmented inputs.
- The experimental results on both simulated and real datasets show that PhaTYP can render better performance than the state-of-the-art methods. In addition,

we applied PhaTYP to predict lifestyles for phage contigs assembled from human neonates' gut data, demonstrating that PhaTYP can help researchers obtain new knowledge on phage composition in microbiome.

Data availability

Funding

References

1. McGrath S, van Sinderen, *et al. Bacteriophage: Genetics and Molecular Biology*. Wymondham, UK: Caister Academic Press, 2007.
2. Zhong Z-P, Tian F, Simon Roux M, *et al.* Glacier ice archives nearly 15,000-year-old microbes and phages. *Microbiome* 2021;**9**(1):1–23.
3. Nishimura Y, Watai H, Honda T, *et al.* Environmental viral genomes shed new light on virus-host interactions in the ocean. *Msphere* 2017;**2**(2):e00359–16.
4. Moineau S. Applications of phage resistance in lactic acid bacteria. *Lactic Acid Bact* 1999;**76**(1–4):377–82.
5. Brüssow H, Desiere F. Comparative phage genomics and the evolution of siphoviridae: insights from dairy phages. *Mol Microbiol* 2001;**39**(2):213–23.
6. Azimi T, Mosadegh M, Nasiri MJ, *et al.* Phage therapy as a renewed therapeutic approach to mycobacterial infections: a comprehensive review. *Infect Drug Resist* 2019;**12**:2943.
7. Loc-Carrillo C, Abedon ST. Pros and cons of phage therapy. *Bacteriophage* 2011;**1**(2):111–4.
8. Wang L-F, Meng Y. Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Curr Drug Targets* 2004;**5**(1):1–15.
9. Bazan J, Całkosiński I, Gamian A. Phage display-A powerful technique for immunotherapy: 1. Introduction and potential of therapeutic applications. *Hum Vaccin Immunother* 2012;**8**(12):1817–28.
10. Shkoporov AN, Khokhlova EV, Brian Fitzgerald C, *et al.* ϕCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun* 2018;**9**(1):1–8.
11. Mirzaei MK, Maurice CF. Ménage à trois in the human gut: interactions between host, bacteria and phages. *Nat Rev Microbiol* 2017;**15**(7):397–408.
12. Clarke KJ. Virus particle production in lysogenic bacteria exposed to protozoan grazing. *FEMS Microbiol Lett* 1998;**166**(2):177–80.
13. Clark DW, Meyer H-P, Leist C, *et al.* Effects of growth medium on phage production and induction in *Escherichia coli* k-12 lambda lysogens. *J Biotechnol* 1986;**3**(5–6):271–80.
14. Housby JN, Mann NH. Phage therapy. *Drug Discov Today* 2009;**14**(11–12):536–40.
15. Brives C, Pourraz J. Phage therapy as a potential solution in the fight against AMR: obstacles and possible futures. *Palgrave Commun* 2020;**6**(1):1–11.
16. Menouni R, Hutinet G, Petit M-A, *et al.* Bacterial genome remodeling through bacteriophage recombination. *FEMS Microbiol Lett* 2015;**362**(1):1–10.
17. Feiner R, Argov T, Rabinovich L, *et al.* A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Microbiol* 2015;**13**(10):641–50.
18. Howard-Varona C, Hargreaves KR, Abedon ST, *et al.* Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J* 2017;**11**(7):1511–20.
19. McNair K, Bailey BA, Edwards RA. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* 2012;**28**(5):614–8.
20. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, *et al.* Massive expansion of human gut bacteriophage diversity. *Cell* 2021;**184**(4):1098–109.
21. Hargreaves KR, Clokie MRJ. *Clostridium difficile* phages: still difficult? *Front Microbiol* 2014;**5**:184.
22. Xiong X, Zhang H-M, Ting-Ting W, *et al.* Titer dynamic analysis of d29 within MTB-infected macrophages and effect on immune function of macrophages. *Exp Lung Res* 2014;**40**(2):86–98.
23. Carrigy NB, Larsen SE, Reese V, *et al.* Prophylaxis of *Mycobacterium tuberculosis* h37rv infection in a preclinical mouse model via inhalation of nebulized bacteriophage d29. *Antimicrob Agents Chemother* 2019;**63**(12):e00871–19.
24. Sepulveda BP, Redgwell T, Rihtman B, *et al.* Marine phage genomics: the tip of the iceberg. *FEMS Microbiol Lett* 2016;**363**(15):fnw158.
25. Shufang W, Fang Z, Tan J, *et al.* Deephage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. *GigaScience* 2021;**10**(9):giab056.
26. Roux S, Páez-Espino D, Chen I-MA, *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res* 2020;**49**(D1):D764–75.
27. Edwards RA, McNair K, Faust K, *et al.* Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev* 2016;**40**(2):258–72.
28. Pons JC, Paez-Espino D, Riera G, *et al.* VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. *Bioinformatics* 2021;**37**(13):1805–13.
29. Jang HB, Bolduc B, Zablocki O, *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* 2019;**37**(6):632–9.
30. Shang J, Jiang J, Sun Y. Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics* 2021;**37**(Supplement_1):i25, 07–33.
31. Shang J, Sun Y. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC Biol* 2021;**19**(1):1–15.
32. Amgarten D, Iha BKV, Piroupo CM, *et al. vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks.* bioRxiv, 2020.
33. Wang W, Ren J, Tang K, *et al.* A network-based integrated framework for predicting virus-prokaryote interactions. *NAR Genomics Bioinformatics* 2020;**2**(2):06.
34. Shang J, Sun Y. CHERRY: a Computational metHod for accuratE pRediction of virus-pRokarYotic interactions using a graph encoder-decoder model. *Brief Bioinform* 2022;**34**(5):bbac182.
35. Shang J, Tang X, Guo R, *et al.* Accurate identification of bacteriophages from metagenomic data using transformer. *Brief Bioinform* 2022;**23**(4):bbac258.

36. Ren J, Ahlgren NA, Yang Young L, *et al.* VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 2017;**5**(1):1–20.

37. Nami Y, Imeni N, Panahi B. Application of machine learning in bacteriophage research. *BMC Microbiol* 2021;**21**(1):1–8.

38. Hockenberry AJ, Wilke CO. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* 2021;**9**: e11396.

39. Emerson JB, Thomas BC, Andrade K, *et al.* Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl Environ Microbiol* 2012;**78**(17):6309–20.

40. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;**7**(10):e1002195.

41. Song K. Classifying the lifestyle of metagenomically-derived phages sequences using alignment-free methods. *Front Microbiol* 2020;**11**:2865.

42. Chibani CM, Farr A, Klama S, *et al.* Classifying the unclassified: a phage classification method. *Viruses* 2019;**11**(2):195.

43. Pfister P, Wasserfallen A, Stettler R, *et al.* Molecular analysis of methanobacterium phage ψm2. *Mol Microbiol*, **30**(2):233–44, 1998.

44. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Adv Neural Inform Process Syst* 2017;**30**:5998–6008.

45. Nambiar A, Heflin M, Liu S, *et al.* Transforming the language of life: transformer neural networks for protein prediction tasks. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, New York, NY, United States: Association for Computing Machinery, 2020, 1–8.

46. Wei D, Xuan Zhao Y, Sun LZ, *et al.* SecProCT: in silico prediction of human secretory proteins based on capsule network and transformer. *Int J Mol Sci* 2021;**22**(16):9054.

47. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**(1):59–60.

48. Enright AJ, Van Dongen, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;**30**(7):1575–84.

49. Hyatt D, Chen G-L, LoCascio PF, *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**(1):1–11.

50. Devlin J, Chang M-W, Lee K, *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.

51. Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality. *Adv Neural Inform Process Syst* 2013;**26**:3111–9.

52. He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Manhattan, New York, NY, United Stated: Institute of Electrical and Electronics Engineers, 2016, 770–8.

53. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint arXiv:1607.06450. 2016.

54. Liang G, Zhao C, Zhang H, *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* 2020;**581**(7809):470–4.

55. Baker DN, Langmead B. Dashing: fast and accurate genomic distances with HyperLogLog. *Genome Biol* 2019;**20**(1):1–12.

56. Dutilh BE, Cassman N, McNair K, *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 2014;**5**:4498.

57. Yutin N, Makarova KS, Gussow AB, *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* 2018;**3**(1): 38–46.

58. Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**(1):1–9.

59. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**(4):357–9.

60. Bushnell B. *Bbmap: A Fast, Accurate, Splice-Aware Aligner*, Technical Report, *Lawrence Berkeley National Lab. (LBNL)*. Berkeley, CA, USA, 2014.

61. Zink RALF, Loessner MARTINJ. Classification of virulent and temperate bacteriophages of listeria spp. on the basis of morphology and protein analysis. *Appl Environ Microbiol* 1992;**58**(1):296–302.

62. Denes T, Vongkamjan K, Ackermann H-W, *et al.* Comparative genomic and morphological analyses of listeria phages isolated from farm environments. *Appl Environ Microbiol* 2014;**80**(15):4616–25.

63. Binns C, Lee MK, Low WY. The long-term public health benefits of breastfeeding. *Asia Pacific J Public Health* 2016;**28**(1):7–14.

64. Allen J, Hector D. Benefits of breastfeeding. *N S W Public Health Bull* 2005;**16**(4):42–6.

65. Liang G, Bushman FD. The human virome: assembly, composition and host interactions. *Nat Rev Microbiol* 2021;**19**(8):514–27.